International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02182-9

Learning Feature Restoration Transformer for Robust Dehazing Visual Object Tracking

Tianyang Xu[1] · Yifan Pan[1] · Zhenhua Feng[2] · Xuefeng Zhu[1] · Chunyang Cheng[1] · Xiao-Jun Wu[1] · Josef Kittler[2]

Received: 15 December 2023 / Accepted: 4 July 2024

Abstract

In recent years, deep-learning-based visual object tracking has obtained promising results. However, a drastic performance drop is observed when transferring a pre-trained model to changing weather conditions, such as hazy imaging scenarios, where the data distribution differs from that of a natural training set. This problem challenges the open-world practical applications of accurate target tracking. In principle, visual tracking performance relies on the discriminative degree of features between the target and its surroundings, rather than the image-level visual quality. To this end, we design a feature restoration transformer that adaptively enhances the representation capability of the extracted visual features for robust tracking in both natural and hazy scenarios. Specifically, a feature restoration transformer is constructed with dedicated self-attention hierarchies for the refinement of potentially contaminated deep feature maps. We endow the feature extraction process with a refinement mechanism typically for hazy imaging scenarios, establishing a tracking system that is robust against foggy videos. In essence, the feature restoration transformer is jointly trained with a Siamese tracking transformer. Intuitively, the supervision for learning discriminative and salient features is facilitated by the entire restoration tracking system. The experimental results obtained on hazy imaging scenarios demonstrate the merits and superiority of the proposed restoration tracking system, with complementary restoration power to image-level dehazing. In addition, consistent advantages of our design can be observed when generalised to different video attributes, demonstrating its capacity to deal with open-world scenarios.

Keywords Visual object tracking · Dehazing system · Siamese tracker · Feature restoration

1 Introduction

Visual object tracking is an omnipresent task in intelligent video analysis systems. The related research is motivated by the demanding downstream practical applications in advanced surveillance, human-computer interaction, medical analysis, smart transportation, etc (Jiao et al., 2021). In general, given the initial state of an object, a tracking system is tasked to estimate the target location in the subsequent video frames. In past decades, with the continuous achievement of the underpinning theory and the tracking methodology, the visual tracking community has witnessed a transition of the research focus from constructing object motion models to object appearance models. In recent years, empowered by the unprecedented availability of a large volume of annotated video data and high-performance

Communicated by Hong Liu.

✉ Tianyang Xu
tianyang.xu@jiangnan.edu.cn

Yifan Pan
6223112023@stu.jiangnan.edu.cn

Zhenhua Feng
z.feng@surrey.ac.uk

Xuefeng Zhu
xuefeng_zhu95@163.com

Chunyang Cheng
7223115009@stu.jiangnan.edu.cn

Xiao-Jun Wu
wu_xiaojun@jiangnan.edu.cn

Josef Kittler
j.kittler@surrey.ac.uk

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China

[2] School of Computer Science and Electronic Engineering and the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK

Published online: 12 July 2024

computing devices, the research community has reported rapid growth of deep neural networks for the tracking task (Bertinetto et al., 2016; Danelljan et al., 2019; Li et al., 2018; Wang et al., 2021). As existing annotated training datasets are collected from daily scenarios, most of the sequences are of high resolution with good illumination conditions. Thereby, training complex convolutional neural networks (CNN) or transformer-based trackers from these datasets has become the mainstream practice in the community. Despite the remarkable advances in the ability to track even targets in the presence of practical challenges, such as deformation, blur, clutter, and occlusion, the transferability of existing tracking approaches to open-world scenarios is still severely limited. In general, in spite of its practical importance, the tracking accuracy and robustness in severe weather situations still fail to meet the requirements of many downstream video applications. Therefore, the manifested necessity of dehazing and deraining visual object tracking inspires our design proposed in this paper.

Among deep-learning-based tracking approaches, the Siamese architecture has been widely explored and studied in recent years. This architecture aims to obtain a target appearance invariant feature space via offline training from paired image patches (template and search region) (Bertinetto et al., 2016; Li et al., 2019). Both foreground awareness and background discrimination are learned in a supervised manner. In particular, classical Siamese architecture highlights the foreground and suppresses the backgrounds via allocating a uni-modal Gaussian heatmap during training. In this setting, after learning the variation contained in several template-search pairs, the appearance invariant feature space can be obtained. As a result, a temporarily changing target can maintain a high degree of consistency of its intrinsic appearance in the feature space, thus being robust to potential appearance variations and background clutters.

Despite the effectiveness of Siamese architecture in learning invariant features against appearance variations, classical Siamese networks suffer from limited localisation granularity. The main reason comes from the simple connections in these Siamese approaches, i.e., a backbone network plus a cross-correlation module (Bertinetto et al., 2016). Drawing on this, the advanced Siamese trackers introduce dedicated functional modules into the basic architecture to achieve fine-grained localisation capacity. Since a visual object tracking problem is a specified target detection in a sequence of frames, many modelling techniques for advanced detector designs have been explored for this task. For instance, to accurately characterise the target appearance, bounding box regression has been introduced in the Siamese paradigm to achieve high-precision target prediction (Li et al., 2018). The regression branch adaptively localises the boundaries of a changing target, providing improved precision against spatio-temporal variations. To alleviate the additional computation consumption for assigning anchors, SiamFC++ further employs an anchor-free index strategy to directly predict the bounding boxes for the pre-assigned spatial points in the obtained representations (Xu et al., 2020b). Besides, to balance the accuracy and stability in predicting the bounding box, centreness is imposed in the regression branch to further improve the tracking robustness (Gao et al., 2020).

More recently, other techniques have also been deployed in constructing advanced trackers in the Siamese paradigm, such as pixel-level mask prediction (Wang et al., 2019), spatio-temporal memory templates (Cao et al., 2022), and multi-task interactions (Xu et al., 2023), resulting in excellent tracking performance on the well-known benchmarks. However, the effectiveness of these offline-trained Siamese approaches relies heavily on the consistency between the training and test data. In practical applications, the above assumption is always challenged by unpredictable data distributions, especially in unconstrained imaging scenarios. In particular, hazy lighting condition is a typical challenging factor, destroying the light propagation from the target surface to the camera sensor.

To guarantee the performance of visual analysis systems in hazy imaging scenarios, existing approaches can be categorised into two main paradigms. The first one is to remove the non-essential appearance from the captured images using image-level restoration techniques. This has been widely studied for image dehazing (Qin et al., 2020; Zhu et al., 2018). The underlying assumption of this paradigm is two-fold: 1) the neighbouring pixels in local regions can provide necessary support information for pixel reconstruction and restoration; 2) the external imaging factor can be modelled with specific patterns using big training data. Such a paradigm has been widely studied in the context of low-level visual tasks, demonstrating impressive success in removing undesirable information from hazy images.

The second approach to addressing the transfer degradation issue involves training a specific supervised model using the data captured in the corresponding hazy imaging scenario. Though guaranteed performance can be achieved by collecting the data and training a network in specific imaging scenarios, it is always time-consuming to pursue such an approach. A training-friendly dataset requires carefully assigned sample distribution, with diverse categories, scales, illuminations, and motion types. Compared to the cost of collecting video clips, more resource is allocated to labelling the ground truth. More importantly, it is difficult to collect naturally hazy data. Therefore, it is preferable to construct a general tracking system that can adaptively enhance the visual clues derived from hazy imaging scenarios.

In this paper, we propose a novel Siamese-based dehazing tracking approach enabled by a general feature restoration module (FRT). To improve the model transferability, we apply an innovative restoration process at the feature level.

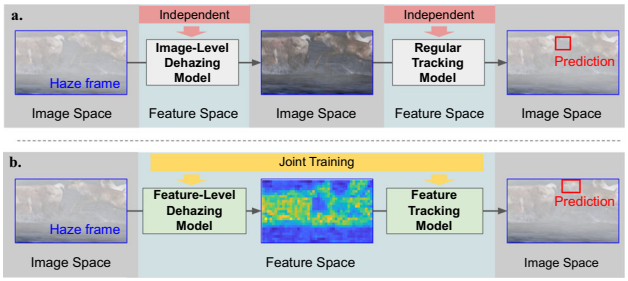

Fig. 1 Comparison of existing dehazing tracking approach and our formulation. Rather than performing image-level dehazing and regular tracking separately, we propose to achieve joint feature-level dehazing and tracking, unifying the processing in the same feature space

As shown in Fig. 1, compared with the existing image-level dehazing restoration techniques, the proposed restoration module is designed to refine the feature representations rather than the input pixels. The FRT module aims to recover the discriminative feature patterns that can distinguish the object from the surroundings even in hazy scenarios. Different from the image-level dehazing approaches that can only process the dehazy data, our FRT enables joint natural and hazy video tracking.

Different from existing transformer structures, to enable its haze restoration power, we introduce restoration prompts and the local attention mechanism, correspondingly. The restoration prompts aim to guide the restoration processing within the encoder and decoder layers. Besides, it is not necessary to access the long-range dependencies in the restoration stage, we design a local attention mechanism to accelerate the token aggregation. The information in the transformer layers (Dosovitskiy et al., 2020) is hierarchically accumulated in a U-shape manner to extract and restore the underlying salient features (Fig. 3). To better reflect the consistency between template-search pairs in the Siamese paradigm, we propose to simultaneously predict the bounding box for both patches, rather than merely focusing on the search patch. As a result, a model of complex dependency between template-search pairs can be gradually obtained by the Siamese tracking transformer. In principle, FRT is jointly trained with the Siamese tracking transformer.

According to the above analysis, to validate network training, hazy video data is necessary. Therefore, rather than manually collecting and annotating hazy videos, we propose to generate hazy video sequences using existing mature haze generation approaches (Zhang et al., 2017; Hong et al., 2020). Specifically, we generate the Haze-COCO dataset for network training. Besides, we evaluate the tracking performance of the proposed feature restoration transformer tracker (FRTT) on synthesised hazy video datasets, i.e., Haze-OTB2015, Haze-TC128, and Haze-VOT2018. The experimental results demonstrate the merits of performing feature restoration for visual tracking in hazy scenarios.

To summarise, the proposed approach has three main innovations:

- A feature restoration transformer is designed to refine the hazy visual features to improve the target discrimination in visual object tracking, thus providing a robust feature enhancement for foggy imaging scenarios.
- A novel multi-task Siamese transformer is proposed to achieve inter- and intra-interactions of each template-search pair, improving the supervision signal with the support from the template target self-perception.
- The proposed FRTT achieves robust dehazing visual tracking performance in challenging foggy videos, providing a novel dehazing solution for downstream vision tasks, which is generalised to a deraining solution with consistent performance.

The remainder of this paper is organised as follows. In Sect. 2, we introduce relevant development of recent tracking paradigms and dehazing studies. Our joint dehazing and tracking network is presented in Sect. 3. The experimental results and analysis are reported in Sect. 4. In Sect. 5, we summarise our work into conclusions.

2 Related Work

For a detailed review of the existing tracking frameworks, a reader can refer to recent surveys (Wu et al., 2013; Li et al., 2016; Wu et al., 2015). In this paper, we briefly discuss relevant studies that formulate the tracking paradigms and inspire our designs. The following review includes advanced tracking approaches and dehazing formulations.

2.1 Discriminative and Siamese Tracking Approaches

Recent progress in visual tracking focuses on training online discriminative filters and offline Siamese networks. The online discriminative paradigm follows the traditional learning-detection-updating manner to obtain a list of dedicated discriminative filters (Danelljan et al., 2019; Bhat et al., 2019). Thus, the target can be highlighted by performing a correlation between the filters and the corresponding feature representations. The learning stage is to optimise a bank of correlation filters that fit the desired output with regularisations. In terms of constraining the distribution of the obtained filters, several studies have explored spatial weighting strategies (Danelljan et al., 2017), channel selection embeddings (Xu et al., 2019, 2021), and multi-feature fusion schemes (Danelljan et al., 2019; Xu et al., 2020).

Besides the online learning paradigm, to exploit the underlying appearance correspondence in the annotated video

datasets, offline pair-wise matching is accomplished by deep neural networks based on the Siamese architecture. In particular, the Siamese structure inputs an image pair, i.e., a template and a search region, to obtain feature representations by a backbone network and several adjustment modules, where the target regions within the template and search region are of salient similarity, and free of nuisance variables confounding the appearance variations. After obtaining the backbone feature embedding, SINT (Tao et al., 2016) calculates the scores of each template-candidate pair, where the candidates are cropped from the search region. Then, the candidate corresponding to the maximal score defines the final target location. To avoid complicated calculations for the sampled template-search pairs, GOTURN (Held et al., 2016) directly predicts the bounding box location using stacked fully connected (FC) layers. In spite of the simple design of GOTURN, the FC layers restrict the input resolution of the network. To enable flexible input sizes, the cross-correlation between template and search is directly used to obtain the response map in SiameseFC (Bertinetto et al., 2016), which further boosts the merit of the Siamese tracking paradigm.

The above Siamese approaches construct an end-to-end learning network to infer the centre location of the target in the search region, for a given template. Besides the module to localise the target centre, precise allocation of the target scale (width and height) is significant for accurate tracking. Therefore, existing studies have concentrated on incorporating bounding box regression designs in the Siamese structures. Specifically, SiamRPN proposes to employ the Region Proposal Network (RPN) to perform simultaneous classification and regression and each anchor (Li et al., 2018). A similar multi-branch structure has been developed with anchor-free mode to alleviate the computational burden from anchor assignment (Xu et al., 2020b). Specifically, the predefined anchors are replaced by grid points, such that the number of candidates decreases k times, where k denotes the scale numbers. Besides constructing multi-task structures and formulating anchor-free models, ATOM performs IOU maximisation for the template-candidate pair, achieving fine-grained overlap precision (Danelljan et al., 2019). In addition, SiamRPN++ (Li et al., 2019) has been proposed to effectively fuse the features obtained by different Siamese layers.

2.2 Transformer Tracking Approaches

In the recent two years, the mainstream modelling techniques in many computer vision tasks have shifted from CNNs to transformer models. The global perception ability exhibited by the self-attention mechanism enables comprehensive relationship mapping within the input tokens. In visual object tracking, the transformer blocks are first introduced in TransT Chen et al. (2021). The cross-attention layers compute the cross-interactions between a template and a search, hierarchically injecting the intrinsic target information from the template into the search region (Wang et al., 2021). Such cross-interaction operations obtain improved fusion results than the widely used simple correlation (Li et al., 2018) and graph mapping (Gao et al., 2021). To extend the cross-attention in a more general perspective, both inter- and intra-interaction within the template and search region are emphasised in Stark (Yan et al., 2021). Then, the feature maps are directly concatenated and flattened for self-attention training. Besides, existing studies also explore the potential of transformer trackers in terms of hierarchical feature fusion (Cao et al., 2021; Kang et al., 2023), refined attention modules (Xu et al., 2023; Gao et al., 2022), and pure Transformer architectures (Cui et al., 2022a; Li et al., 2023).

Although the developed tracking approaches have lifted up the benchmark record, the performance degradation during the online tracking stage in hazy videos is underestimated. In principle, existing tracking processes are predefined by specified parameterised models regardless of the input video imaging conditions. These models always assume consistent distributions of the target appearance and lighting scenarios between the training and test datasets. To bridge this gap, we argue that the tracking process should benefit from appearance adaptation in hazy situations. This is possible only if a reliable restoration mechanism for each hazy input can be provided so that an effective appearance correction can be applied.

2.3 Dehazing Formulations

In low-level dehazing approaches, physical scattering models were first introduced to directly formulate the inverse hazy imaging scenario problem using traditional prior-based algorithms. In this paradigm, DCP (He et al., 2010) proposes to use a prior clue to estimate the transmission map, explicitly modelling the difference between the image object and atmospheric light. Similar formulations are suggested for colour attenuation (Zhu et al., 2014) and non-local prior (Berman et al., 2016). The effectiveness of these approaches is determined by the fitting degree of the transmission map to the target scenarios. However, a complex transmission map can not be predicted well for a given hazy image, using handcrafted models, especially in unconstrained scenarios.

To benefit from large image datasets, deep learning has been naturally applied to perform image dehazing in recent years. DehazeNet (Cai et al., 2016) is the seminal approach that constructs an end-to-end network to estimate the medium transmission map. The local modelling and nonlinear activation of CNNs further improve the transmission capacity as compared to the traditional approaches. A similar structure was further introduced (Ren et al., 2016) to support multi-scale perception, achieving a transmission map refinement.

To alleviate the accumulated artifacts, advanced deep dehazing methods focus on directly generating restored images, rather than the transmission map. To this end, a feature fusion attention mechanism was proposed to obtain the desired output, with ℓ_1 loss to supervise the network training (Qin et al., 2020). Besides, positive samples were used (Hong et al., 2020) through teacher-student supervision to achieve knowledge transfer.

Though advanced performance has been achieved for these image-level dehazing approaches, we argue that image-level dehazing is not the only solution for visual recognition tasks in hazy scenarios. In principle, we propose to perform joint feature-level dehazing and tracking in an end-to-end learning network. Specifically, it has been demonstrated that feature representations are significant in delivering promising visual tracking performance (Wang et al., 2015). This suggests that for downstream visual tasks, such as visual tracking, it is more necessary to restore the haze-free descriptors rather than directly recover the photo-realistic image pixels.

3 Feature Restoration Transformer Tracker (FRTT)

In this section, we introduce our feature restoration transformer tracking approach FRTT in detail. As shown in Fig. 2, we input a pair of hazy images into FRTT, i.e., hazy target template and hazy search region. The area of the template is 4 times the target scale, while the area of the search region is 25 times the target. Both the template and search region are squared cropped from two randomly selected frames within a video sequence. They are passed into a shared pre-trained CNN backbone network to obtain the feature representations, and then flattened and concatenated into ordered tokens $X = \{X^T, X^I\}$, where the superscripts T and I denote template and search region, respectively. The proposed feature restoration transformer performs feature token refinement, supervised by the tokens extracted from clear-view image pairs. Once the restored tokens $\hat{X}$ are obtained, a template-

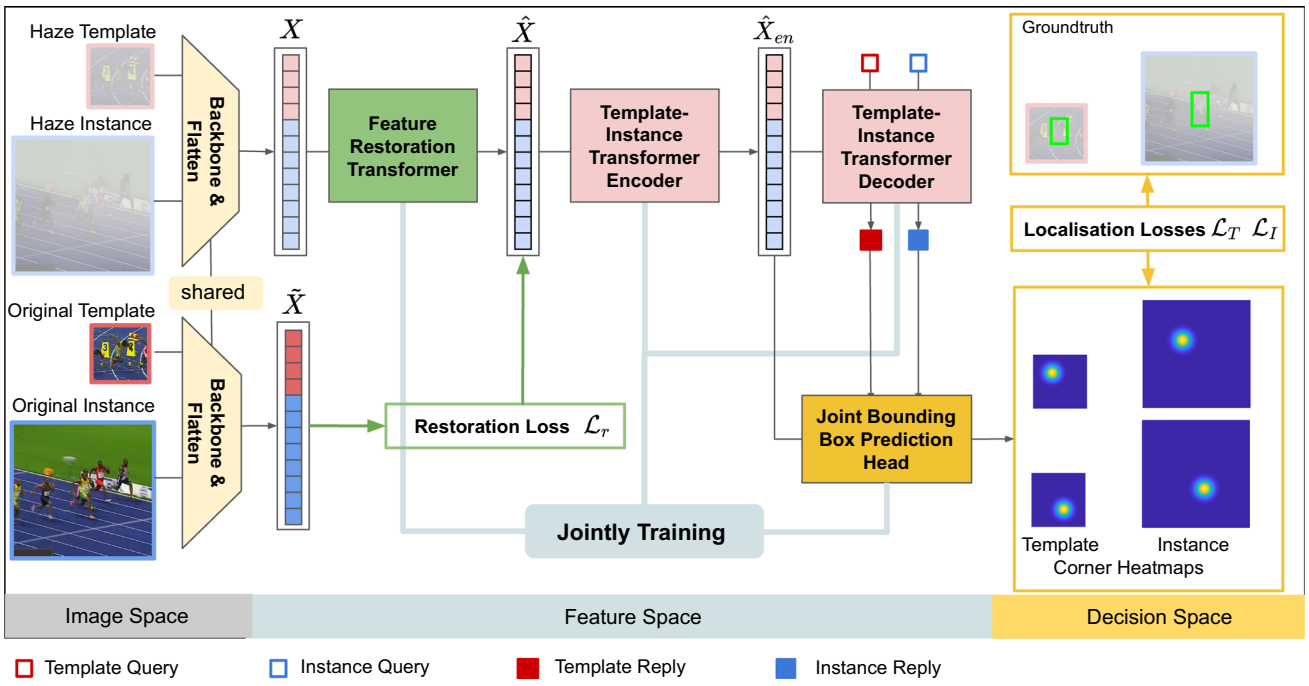

Fig. 2 An overview of our FRTT that has the following components: backbone feature extractor, feature restoration transformer, template-search transformer encoder, template-search transformer decoder, and joint bounding box prediction head. The backbone network is used to extract and adjust hazy input features X, as introduced in Sec. 3.1. The feature restoration transformer is designed to refine the hazy visual features from X to $\hat{X}$ to enhance their discriminative power, as presented in Sec. 3.2. The template-search encoder-decoder is constructed to achieve a spatio-temporal fusion, obtaining $\hat{X}_{en}$ which captures long-term dependency interactions to highlight the changing target, as introduced in Sec. 3.3. The joint bounding box prediction head is constructed to perform a joint corner heatmaps prediction, as described in Sec. 3.4

search transformer encoder is used to fuse the inter- and intra-relations among $\hat{X}^T$ and $\hat{X}^I$. Therefore, the template reference can be absorbed by the search tokens. Two query entities and the encoder output $\hat{X}_{en}$ are then input into the decoder to obtain the corresponding reply tokens. For the final prediction heads, we simultaneously predict the corner heatmaps for both the template and search region.

3.1 Backbone Network

In the proposed method, both ResNet-50 (He et al., 2016) and CvT (Wu et al., 2021) are tested, separately, as the backbone feature extractor. Therefore, both convolution-based and transformer-based backbones are involved to reflect the consistent capacity of our design. We use the weights pretrained on ImageNet-1K for network initialisation. In general, other backbone networks can also be considered to substitute ResNet-50 or CvT to balance the effectiveness and efficiency in practice, which has been widely explored in edge device-based applications, such as UAV tracking problems (Cao et al., 2022). Specifically, the input hazy template and search region are of size $H^T \times W^T \times 3$ and $H^I \times W^I \times 3$, respectively. They are synthesised from clear-view images by HazeRD (Zhang et al., 2017), as no annotated hazy video dataset is available. The obtained feature maps are of $\frac{H^T}{s} \times \frac{W^T}{s} \times C$ and $\frac{H^I}{s} \times \frac{W^I}{s} \times C$, where s is the total stride. The subsequent one 1×1 conv layer is applied to reduce the channel number from C to D. Similar to ViT (Dosovitskiy et al., 2020), we flatten the spatial dimensions of the hazy image feature maps and concatenate them together to obtain the feature tokens $X \in \mathbb{R}^{\frac{W^T H^T + W^I H^I}{s^2} \times D}$. In the training stage, the clear-view images are also considered to extract the corresponding feature tokens $\tilde{X}$, for the use of a supervision signal in the restoration stage. To link the spatial layout among input tokens, positional embeddings are assigned to these tokens with a fixed cosine manner (Dosovitskiy et al., 2020).

3.2 Feature Restoration Network

Note that the quality of feature representation is the most essential factor for tracking. However, haze directly contaminates the visibility and clarity of the raw pixels and degrades the quality of image features. Though existing image-level dehazing approaches have achieved significant progress, we argue that it is not necessary to recover the raw pixels for the visual tracking task. In principle, a high-performance tracker focuses on the discriminative clues between the target and surroundings, rather than reconstructing and enhancing the raw pixels. Therefore, we propose performing feature restoration for the contaminated backbone features. Given the feature tokens obtained from both hazy and clean views, X and $\tilde{X}$, the target of our feature restoration network is to

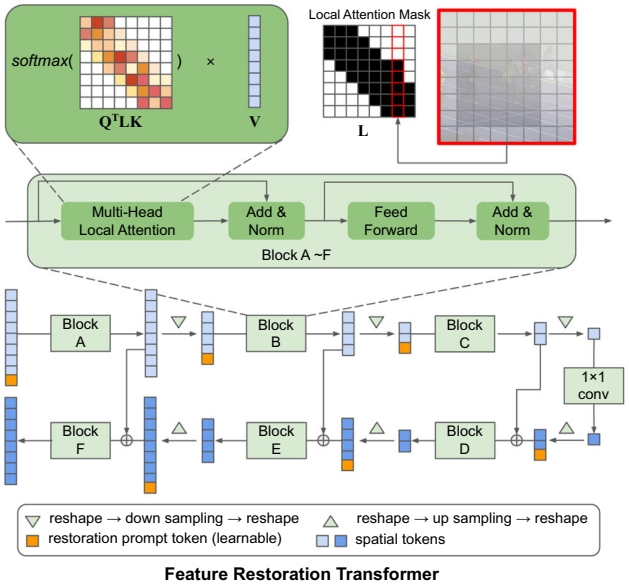

Fig. 3 The architecture of the proposed feature restoration transformer

recover the discriminative and salient feature representations from X. As shown in Fig. 3, stacked feature restoration transformer blocks are utilised with skip connections to gradually refine the involved feature tokens. Specifically, 3 downsampling operators are inserted after Block A $\sim$ C in the upper branch, compressing the token number to $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$, respectively. While the 1×1 conv layer is used to perform channel adjustment. In the bottom branch, Block D $\sim$ F are stacked to hierarchically reconstruct the features, where skip connections are used to concatenate the corresponding output of the upper blocks. For each block, besides the feature tokens, a learnable restoration prompt is involved to guide the interaction in the local attention layer. In the local attention layer, we propose a multi-head local mechanism to perform token interaction rather than the standard self-attention. The local attention is achieved by introducing a local attention mask $\mathbf{L}$ in the multiplication between the queries and keys, where only the neighbouring 5×5 tokens and the prompt tokens are used for each query token. The details of each block are shown at the bottom of Fig. 3. Our local attention emphasises the reconstructive power embedded in the neighbouring region of each spatial region. In particular, standard self-attention uses dense involvement among all the spatial tokens, which is expert in generating long-ranged dependencies for recognition tasks. But our goal is to recover the intrinsic feature representations from hazy input rather than globally perceiving the image content. Therefore, our local attention can highlight such locality and decrease the computation compared to the standard self-attention operation.

Different from the standard vision transformers, our network employs a multi-resolution architecture, formulating

a U-shape restoration manner. To this end, local details are coded and recovered in Block A and Block F, while global patterns are reflected by Block C and Block D. Therefore, multiple granularities are involved in the processing stage, delivering fine-grained restoration performance.

A Standard multi-head attention module with layer normalisation and feed-forward network is employed to capture long-range interactions in each layer. In principle, the proposed feature restoration transformer achieves hazy data representation recovery by incremental disentangling the impact of haze from the contaminated visual features. Having obtained the restored feature tokens $\hat{X}$ and the desired feature tokens $\tilde{X}$, the Charbonnier loss (Charbonnier et al., 1994) is used for image restoration:,

$$\mathcal{L}_r = \sqrt{\|\hat{X} - \tilde{X}\|^2 + \varepsilon^2}, \tag{1}$$

where ε is a constant (empirically set to 10^{-2} in this paper). In terms of using the Charbonnier loss, the underlying reason is that it can balance the requirement of sparse and dense distribution of our reconstruction errors. If the reconstruction error is larger than epsilon, sparse regularisation is dominant. While if the reconstruction error is smaller than epsilon, it highlights dense regularisation. In principle, a feature-level dehazing task is indeed an ill-posed problem, such that a simple dense loss is not suitable, especially for challenging scenarios. Given the above analysis, the Charbonnier loss is suitable for supervising our restoration transformer.

3.3 The Template-Search Transformer Encoder-Decoder

In the template-search matching stage, we first follow the basic structure of Stark (Yan et al., 2021) to perform a direct token fusion within $\hat{X}$ to obtain the encoder output $\hat{X}_{en}$. The template-search encoder consists of N_{en} standard transformer blocks. In the decoder, different from the existing transformer tracking networks that only extract the target clue in the search (Carion et al., 2020; Yan et al., 2021), we simultaneously use two query entities (as input) together with $\hat{X}_{en}$ to obtain the responses for both template and search. Through interaction with the search region, the target state conveyed by the template helps to emphasise the token quality within $\hat{X}_{en}$, providing an additional supervision signal for the feature restoration transformer. The decoder has N_{de} stacked transformer blocks. Based on the proposed transformer network, the target state can reliably be estimated from the feature tokens.

3.4 Joint Bounding Box Prediction Head

The details of the template and search bounding box prediction heads are presented in Fig. 4. We modified the approach in Yan et al. (2021) to simultaneously predict the top-left and down-right corners of the target in the template and search, respectively. In particular, we split the encoder output $\hat{X}_{en}$ into $\hat{X}_{en}^T$ and $\hat{X}_{en}^I$, corresponding to the template and the search features. The encoder output and decoder replies are fused to jointly generate the heat maps of two corners for both the template and search region. After obtaining the corner heat maps, the final bounding box is determined by computing the expectation for each point. Then, the generalised IOU loss and ℓ_1 loss are used for supervising the target localisation loss in the template and the search region:

$$\begin{cases} \mathcal{L}_T = \lambda_{iou}^T \mathcal{L}_{iou}\left(\hat{b}^T, \tilde{b}^T\right) + \lambda_{\ell_1}^T \mathcal{L}_{\ell_1}\left(\hat{b}^T, \tilde{b}^T\right) \\ \mathcal{L}_I = \lambda_{iou}^I \mathcal{L}_{iou}\left(\hat{b}^I, \tilde{b}^I\right) + \lambda_{\ell_1}^T \mathcal{L}_{\ell_1}\left(\hat{b}^I, \tilde{b}^I\right) \end{cases}, \tag{2}$$

where $\hat{b}^T$, $\tilde{b}^T$ are the predicted bounding box and ground truth in the template, $\hat{b}^I$, $\tilde{b}^I$ are the predicted result and ground truth for the search region, λ_{iou}^T, $\lambda_{\ell_1}^T$, λ_{iou}^I, and $\lambda_{\ell_1}^I$ are balancing hyper-parameters.

4 Experimental Results

In this section, we first introduce the experimental settings, including the implementation details evaluation dataset, and evaluation metrics. Then we report the comparison analysis with the state-of-the-art tracking approaches. To validate the merit of performing feature-level dehazing tracking, we compare its performance against the image-level dehazing formulation. Ablation studies are followed to present detailed component analysis.

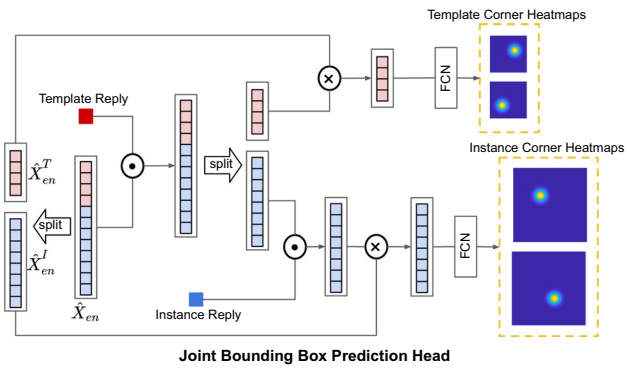

Fig. 4 The architecture of the proposed joint bounding box prediction head

4.1 Experimental Settings

4.1.1 Implementation Details

We implement FRTT with PyTorch on NVIDIA GeForce RTX 2080Ti GPU. To achieve joint training for both the restoration and tracking tasks, we use a mixture of training datasets, including COCO (Lin et al., 2014), LaSOT (Fan et al., 2019), GOT-10k (Huang et al., 2019), TrackingNet (Muller et al., 2018) and the generated Haze-COCO. Haze-COCO is synthesised from the COCO dataset using HazeRD (Zhang et al., 2017). Some generated images are shown in Fig. 5 Here, we only generate the hazy images for COCO as COCO contains still images of 80 object categories and various background scenarios. Except for the data from Haze-COCO, we remove the restoration loss during training and directly copy the backbone features of the original images as the input of the transformer encoder-decoder. The generated Haze-COCO is also involved in training other state-of-the-art tracking networks, enabling fair comparison. The hyperparameters are set as $\lambda_{iou}^{T} = 0.1$, $\lambda_{\ell_1}^{T} = 0.25$, $\lambda_{iou}^{I} = 0.2$, and $\lambda_{\ell_1}^{I} = 0.5$.

To achieve joint feature restoration and tracking, our network is trained in two stages. In the first phase, only the feature restoration transformer is trained with COCO and Haze-COCO, using the ADAMW (Loshchilov & Hutter, 2017) optimiser, for 200 epochs. The learning rate is set to 10^{-4}. The input resolutions of the template and instance are 128×128 and 320×320 pixels, respectively. We use pretrained ResNet-50 and CvT as our backbone network candidates, with the parameters being initialised from ImageNet-1K training. Typically, we name the two versions of our model as FRRT_C and FRRT_V for the two backbones. In the second stage, we train the restoration and tracking modules together with ADAMW for 500 epochs, and 60,000 template-instance pairs are sampled in each epoch. Specifically, the initial learning rate is 10^{-4}, and then decreases to 10^{-5} at the 401 epoch. A gradient norm clipping trick is used, with a predefined threshold of 0.1, to alleviate the impact of gradient explosion. Data augmentation techniques, including random centre and scale jittering, are applied to improve the data diversity. During the inference stage, the initial frame is cropped as the template. We remove the template box head and only output the target state of the instance.

4.1.2 Datasets And Evaluation Metrics

To verify the tracking performance of a method in hazy environments, we generate Haze-OTB2015, Haze-TC128, Haze-VOT2018, Haze-LaSOT, and Haze-GOT-10K from the original OTB2015 (Wu et al., 2015), TC128 (Liang et al., 2015), VOT2018 (Kristan et al., 2018), LaSOT (Fan et al., 2019), and GOT-10K (Huang et al., 2019) using HazeRD. Some typical examples are illustrated in Fig. 5.

For evaluation, we employ the widely used OTB and VOT protocols (Wu et al., 2015; Kristan et al., 2018). The area

Fig. 5 Some examples of the original and generated hazy images from COCO, OTB2015, TC128, and VOT2018

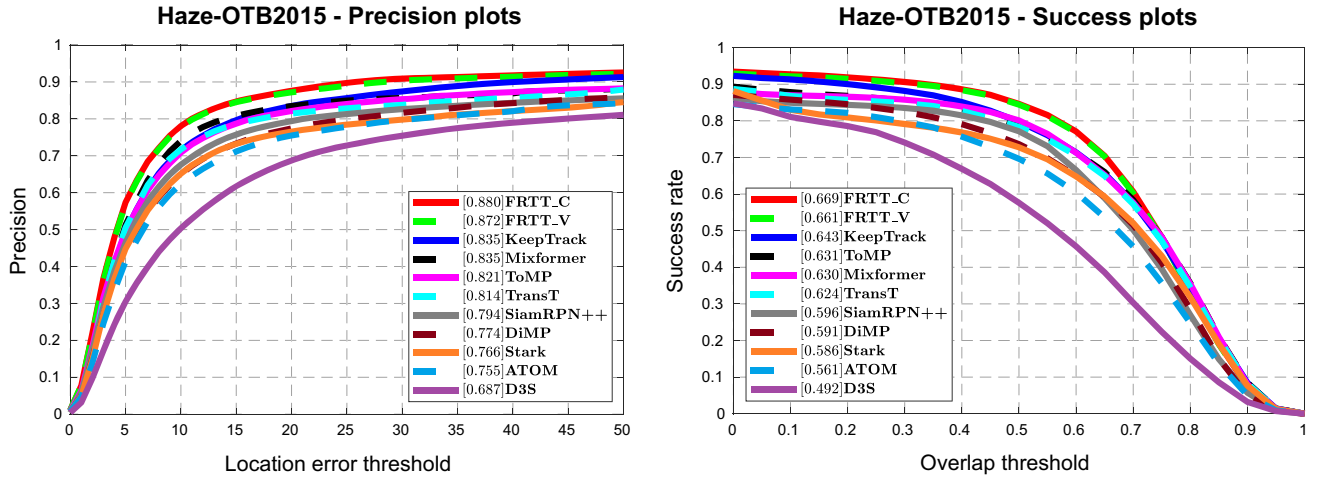

Fig. 6 The experimental results obtained on the Haze-OTB2015 dataset. The precision plots (with the DP score reported in the figure legend) and the success plots (with the AUC score reported in the figure legend) are presented

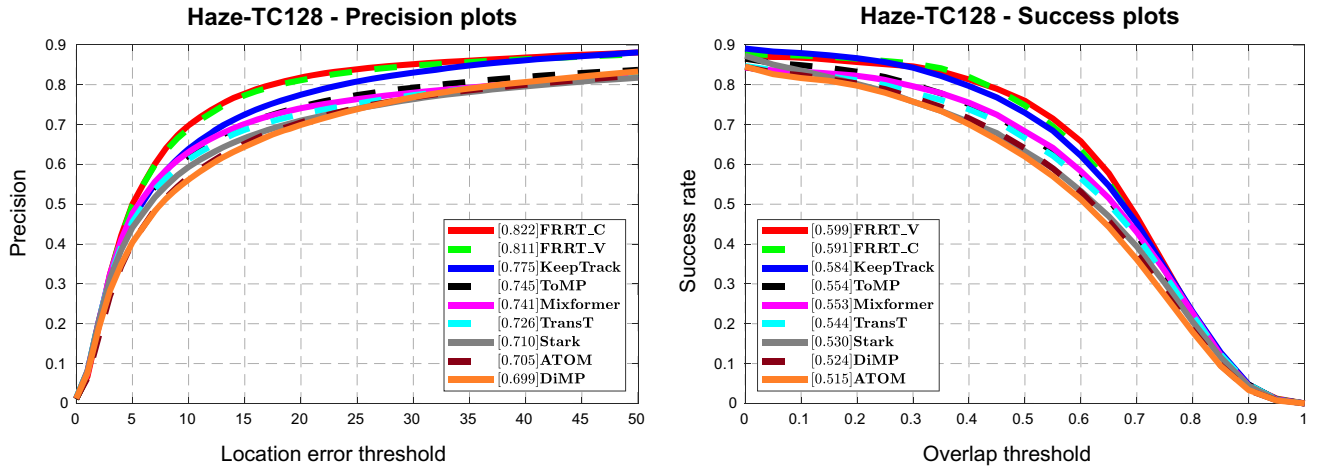

Fig. 7 The experimental results obtained on the Haze-TC128 dataset. The precision plots (with the DP score reported in the figure legend) and the success plots (with the AUC score reported in the figure legend) are presented

Table 1 The tracking results on Haze-VOT2018

	ATOM	DiMP	PrDiMP	SiamRPN^{++}	SiamMask	Stark
EAO	0.271	***0.296***	0.272	0.235	0.223	0.127
A	0.538	0.546	0.563	0.584	0.556	0.587
	TransT	KeepTrack	ToMP	SiamCAR	**FRTT_C**	**FRTT_V**
EAO	0.222	0.170	0.195	0.291	**0.311**	*0.306*
A	**0.601**	***0.593***	**0.601**	0.561	0.571	0.580

Top-performing three results are shown in Bold, Italic and Bolditalic

under curve (AUC) and distance precision (DP) are recorded to generate the success plot and precision plot, respectively, on Haze-OTB2015 and Haze-TC128. Specifically, the success plot reflects the degree of successfully tracked bounding box in terms of overlap performance. A predicted bounding box can be considered as successfully tracked if its overlap with the groundtruth is above a threshold (Wu et al., 2015). In contrast, the precision plot reflects the centre localisation degree. The distance precision (DP) is observed by setting the centre error threshold as 20 pixels. For Haze-VOT2018, we use the expected average overlap (EAO) to comprehensively evaluate the tracking performance (Kristan et al., 2018), which simultaneously measures the bounding box accuracy and robustness against failures. For Haze-LaSOT, precision (PR), normalised precision (NPR), and AUC are used to measure the tracking performance (Fan et al., 2019). For

Table 2 The tracking results on Haze-LaSOT and Haze-GOT-10K

Dataset	Metric	ToMP	PrDiMP	DiMP	ATOM	KeepTrack	Stark	Mixformer	TransT	**FRTT_C**	**FRTT_V**
Haze-LaSOT	**PR**	0.616	0.533	0.504	0.426	0.616	0.598	***0.626***	0.599	**0.658**	*0.644*
	NPR	0.671	0.571	0.541	0.451	0.668	0.649	***0.682***	0.653	**0.722**	*0.713*
	AUC	0.587	0.499	0.476	0.398	0.578	0.575	***0.593***	0.569	**0.614**	*0.605*
Haze-GOT-10K	**AO**	0.605	0.521	0.511	0.426	0.579	0.585	***0.625***	0.588	**0.631**	**0.631**
	SR_0.50	0.702	0.601	0.582	0.483	0.662	0.670	***0.713***	0.671	**0.734**	*0.729*
	SR_0.75	0.509	0.382	0.340	0.269	0.453	0.489	***0.555***	0.496	**0.590**	*0.588*

The top three results are highlighted in Bold, Italic and Bolditalic

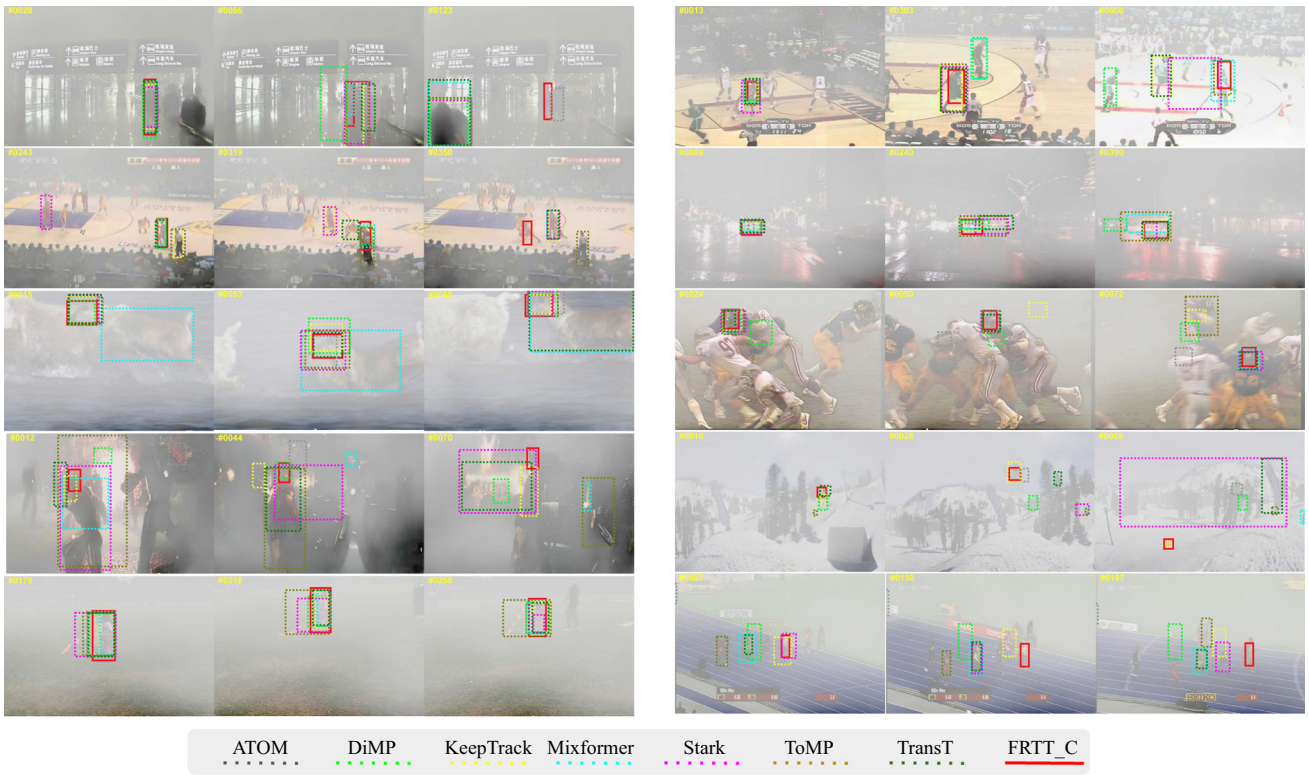

Fig. 8 Illustration of the qualitative tracking results on challenging hazy sequences (the listed videos are *Left*: *Haze-Airport_ce*, *Haze-Basketball_ce2*, *Haze-Deer*, *Haze-Ironman*, *Haze-Baby_ce*, *Right*: *Haze-Basketball_ce1*, *Haze-CarDark*, *Haze-Football1*, *Haze-Skiing*, *Haze-Bolt*, from the Haze-TC128 dataset). The colour bounding boxes are the corresponding predictions of ATOM, DiMP, KeepTrack, Mixformer, Stark, ToMP, TransT, and the proposed FRTT tracker

Haze-GOT-10K, its online evaluation server outputs the test performance in terms of average overlap (AO), the success rates with two thresholds ($SR_{0.50}$ and $SR_{0.75}$) (Huang et al., 2019).

4.2 Comparison With Other Tracking Algorithms

Haze-OTB2015: There are 100 videos in the Haze-OTB2015 dataset. Besides the hazy lighting conditions, the videos also contain other challenging factors, reflecting diverse visual effects for practical variations. As haze is additionally imposed on the original OTB2015 images, it is much more difficult to successfully track the target on Haze-OTB2015. The average video length of Haze-OTB2015 is 591 frames. The experimental results are reported in Fig. 6 using the precision and success plots.

Among all the advanced DCF and Siamese approaches, our FRTT design obtains the top performance, in both precision and success plots. Compared to the third best, KeepTrack, FRTT_C improves the DP from 83.5% to 88.0%, and the AUC from 64.3% to 66.9%, demonstrating its advantages in robust tracking under challenging hazy scenarios. FRTT_V achieves a similar performance compared with FRTT_C. We attribute the slight advantage of FRTT_C to the

Table 3 A relationship of image-level dehazing and our feature-level dehazing trackers on Haze-OTB2015 and Haze-TC128

		SiamRPN++ (Li et al., 2019)	SiamCAR (Gao et al., 2020)	SiamBAN (Chen et al., 2020)	SiamMask (Wang et al., 2019)	ATOM (Danelljan et al., 2019)	DiMP (Bhat et al., 2019)	TransT (Chen et al., 2021)	Mixformer (Cui et al., 2022b)	Stark (Yan et al., 2021)	KeepTrack (Mayer et al., 2021)	ToMP (Mayer et al., 2022)	**FRTT_C**	**FRTT_V**
Haze-OTB2015	**DP w/o ILD**	0.794	0.820	0.793	0.762	0.755	0.774	0.814	***0.835***	0.766	***0.835***	0.821	**0.880**	*0.872*
	DP w ILD	0.800	0.847	0.834	0.790	0.783	0.804	0.824	0.817	0.773	***0.850***	0.832	**0.912**	*0.908*
	DP Δ	**+0.006**	**+0.027**	**+0.041**	**+0.028**	**+0.028**	**+0.030**	**+0.010**	−0.018	**+0.007**	**+0.015**	**+0.011**	**+0.032**	**+0.036**
	AUC w/o ILD	0.596	0.618	0.590	0.572	0.561	0.591	0.624	0.630	0.586	***0.643***	0.631	**0.669**	*0.661*
	AUC w ILD	0.604	0.640	0.630	0.592	0.586	0.612	0.628	0.626	0.593	***0.651***	0.637	**0.692**	*0.680*
	AUC Δ	**+0.008**	**+0.022**	**+0.040**	**+0.020**	**+0.025**	**+0.021**	**+0.004**	−0.004	**+0.007**	**+0.008**	**+0.006**	**+0.023**	**+0.019**
Haze-TC128	**DP w/o ILD**	0.687	0.710	0.683	0.680	0.705	0.699	0.726	0.741	0.710	***0.775***	0.745	**0.822**	*0.817*
	DP w ILD	0.723	0.720	0.708	0.664	0.719	0.731	0.733	0.776	0.719	***0.777***	0.772	**0.824**	*0.815*
	DP Δ	**+0.036**	**+0.010**	**+0.025**	−0.016	**+0.014**	**+0.032**	**+0.007**	**+0.035**	**+0.009**	**+0.002**	**+0.027**	**+0.002**	**+0.002**
	AUC w/o ILD	0.505	0.524	0.496	0.493	0.515	0.524	0.544	0.553	0.530	***0.584***	0.554	*0.591*	**0.599**
	AUC w ILD	0.531	0.535	0.518	0.486	0.528	0.550	0.546	0.574	0.534	***0.590***	0.566	*0.603*	**0.605**
	AUC Δ	**+0.026**	**+0.011**	**+0.022**	−0.007	**+0.013**	**+0.026**	**+0.002**	**+0.021**	**+0.004**	**+0.006**	**+0.012**	**+0.012**	**+0.006**
Haze-VOT2018	**EAO w/o ILD**	0.235	0.291	0.233	0.223	0.271	***0.296***	0.222	0.164	0.127	0.170	0.195	**0.311**	*0.306*
	EAO w ILD	0.255	***0.319***	0.277	0.261	0.303	**0.342**	0.225	0.169	0.133	0.172	0.206	*0.322*	0.310
	EAO Δ	**+0.020**	**+0.028**	**+0.044**	**+0.038**	**+0.032**	**+0.046**	**+0.003**	**+0.005**	**+0.006**	**+0.002**	**+0.011**	**+0.011**	**+0.004**

The top three results are highlighted in Bold, Italic and Bolditalic

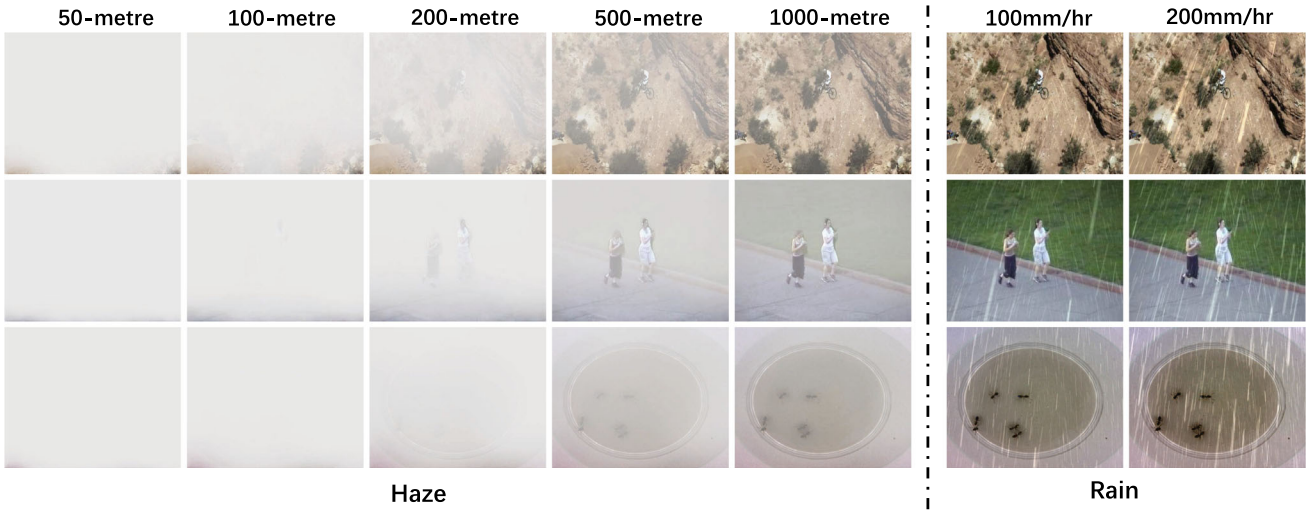

Fig. 9 Visualisation of the haze and rainy samples, with different contamination levels

fact that convolutions are more expert in preserving local relevance than transformers, which exhibit more reconstruction power in our restoration transformer. In particular, SiamMask performs poorly on this benchmark. The underlying reason is that SiamMask performs segmentation to localise the target. However, predicting a precise target mask is too difficult for hazy input as shown in Fig. 5. The results illustrate the merits of performing feature restoration proposed by our approach for hazy videos.

Haze-TC128: The Haze-TC128 dataset contains 128 video sequences. The original TC128 dataset focuses on exploring the potential of colour clues for advanced visual tracking systems and the additional added hazy content challenges the representative power and increases the tracking difficulty. The average sequence length of Haze-TC128 is 429, belonging to standard short-term tracking datasets. We report the experimental results in Fig. 7 using the precision and success plots.

Compared to the advanced end-to-end deep trackers, the FRTT series achieves the top performance in both plots. Consistent promising results of our approach can be observed on Haze-TC128 as on Haze-OTB2015. Specifically, compared to Stark, which contains similar components in modelling the appearance network as FRRT, our FRRT_C outperforms Stark by 11.1% and 6.1% in terms of DP and AUC. As the difficulty of Haze-TC128 is increased than Haze-OTB2015, FRRT_V performs better than FRRT_C in terms of overlap metric (AUC), reflecting the relative merit of long-range dependencies in localising challenging sequences. The above comparison verifies the effectiveness and superiority of performing feature-level dehazing in visual object tracking.

Haze-VOT2018: The Haze-VOT2018 dataset contains 60 challenging video sequences for single object tracking. Following the standard protocol, we report the detailed results in Table 1. Consistent advantages of FRRT_C and FRRT_V can be obtained on Haze-VOT2018, achieving 0.311 and 0.306 in terms of EAO. Specifically, without performing restoration, the transformer-based Stark suffers the largest performance drop, which indicates that the particular imaging challenge cannot be directly addressed by simply constructing powerful transformer trackers. As self-attention amplifies the similar patterns while suppressing the dissimilar ones during token interaction, haze contaminates the token embeddings thus drastically weakening the discrimination.

Haze-LaSOT: The Haze-LaSOT validation set contains 280 long-term video sequences. We report the experimental results in Table 2. Our FRTT_C achieves the best performance among all the involved trackers, obtaining 0.772 and 0.614 in terms of NPR and AUC. Compared to the advanced KeepTrack, both FRTT_C and FRTT_V exhibit more than 2% advantages in terms of PR, NPR, and AUC. In particular, the average length of Haze-LaSOT is more than 1500 frames per sequence, such long-term property verifies the consistent merit of our FRTT.

Haze-GOT-10K: The Haze-GOT-10K test set contains 180 challenging video sequences with real-world moving objects. Table 2 lists the tracking results. Similar to the performance on Haze-LaSOT, our FRTT_C and FRTT_V surpass other advanced trackers.

Besides quantitative comparisons, we also deliver the qualitative performance results in Fig. 8. As in sequences *Haze-Airport_ce* and *Haze-Baby_ce*, their original non-hazy videos are less challenging, and all the advanced trackers can successfully track the target. While in the hazy scenario, our FRRT achieves robust tracking performance compared to the failures produced by other trackers. For other more challenging sequences (*Haze-Ironman*, *Haze-Skiing*), our FRRT can still maintain high-performing predictions, thanks to the

feature restoration capability. The above visualisation comparison reflects a consistent merit of the proposed design in dealing with hazy visual tracking issues.

4.3 Image-Level Dehazing + Tracking v.s. Feature-Level Dehazing Tracking

The entire motivation of our design is constructed on the belief that feature-level dehazing tracking outperforms the image-level dehazing + tracking formulations. Therefore, we further report the tracking performance of image-level dehazing + tracking (ILD+T) and feature-level dehazing tracking (FLDT) approaches. In particular, DehazeFormer (Song et al., 2023) is used to perform image-level dehazing task for the test datasets, i.e., Haze-OTB2015, Haze-TC128, and Haze-VOT2018. After obtaining the dehazed image sequences, we repeat the test process for all the involved trackers. Detailed results are reported in Table 3. For Haze-OTB2015, except Mixformer, all the trackers improve the tracking performance after using image-level pre-process, with the DP gains ranging from 0.6% to 4.1%, and AUC gains ranging from 0.4% to 4.0%. Among all the compared trackers, the best-performing tracker with image-level dehazing + tracking is KeepTrack, which achieves 85.0% DP and 65.1% AUC. While our FRRT method (88.0% DP and 66.9% AUC) still performs better than KeepTrack with image-level dehazing + tracking, demonstrating the merit of performing feature-level dehazing tracking in the proposed method. In addition, FRRT further improves the performance to 91.2% DP and 69.2% AUC after using both image- and feature-level dehazing.

We observe similar results on Haze-TC128. The top-performing tracker after image-level dehazing + tracking is DiMP, with 0.342 EAO, which is better than FRRT. The underlying reason is that VOT exhibits severe short-term visual challenges, DiMP quipped with online updated filters presents a better solution than the approaches using a fixed template. The sequences with image-level dehazing + tracking are less challenging, which can unveil the online modelling capacity of DiMP. Besides the perspective of dehazing power, the forward pass of image-level dehazing (DehazeFormer-M) costs 77.29ms while our feature restoration transformer only needs 6.35ms in our implementation, demonstrating the efficiency of our solution. According to the above analysis, we can conclude the merit of our feature-level dehazing tracking design and the complementary effect between image-level + tracking and feature-level dehazing tracking.

4.4 Dehazing Tracking with Different Challenging Levels

To further analyse the restoration power of our designed approach, we conduct experiments in terms of both hazy and rainy scenarios with different degrees. Illustration of the generated hazy and rainy samples is presented in Fig. 9. As listed in Table 4, different levels of visibility are considered the dehazing ability of our approach and other competitors. The degree denotes the visibility distance (metre), which is obtained by HazeRD. Consistent with the default haze degree (500-metre visibility), our FRTT achieves the best performance in all the possible degrees, ranging from 50 ms to 1000 ms. Similarly, the gap between our two versions, FRTT_C and FRTT_V, is also marginal, where FRTT_C performs slightly better than FRTT_V in most situations. Though outperforming other trackers, our approach cannot well handle the heavy haze, e.g., 50-metre visibility, with the DP and AUC below 0.2. Among the involved advanced trackers, it is interesting that ToMP exhibits a relative advantage in the heavy haze situations. We attribute this as ToMP can capture global information with extremely low inductive bias and thus learn more powerful predictions of the target model, which is quite suitable under extremely hazy scenarios. Given the above analysis, we verified the merit of our restoration design in terms of different haze levels.

Besides haze, in practical scenarios, other challenging weathers also impede the applications of the normally well-trained trackers. Therefore, we perform experiments to verify the generation ability of our approach against the rainy weather. As reported in Table 5, we list the results of tracking performance under 100 mm/hr and 200 mm/hr rainfall situations. Different from the hazy images, rainy images receive less negative impact in terms of tracking performance drop. Most of the trackers can achieve the DP and AUC above 0.8 and 0.6, which already an acceptable results in challenging scenarios. In principle, the absolute values of our FRTT performance are also much higher than those in the hazy situations, but it cannot outperform other advanced trackers. After carefully compare the contamination added by haze and rain, we found that haze also destroys the holistic appearance of an image, while rain only affects a limited number of spatial regions. Therefore, a certain degree of the original appearance can well support the discrimination calculation during the tracking process, which decreases the negative impact of the images with rains. In addition, according to the above analysis, specific designs should be proposed to restore the rainy data, and thus further boost the derainy tracking performance.

Beyond challenging weather, we further provide the tracking performance of our tracking approach on the standard LaSOT, to evaluate its basic capacity for non-hazy videos. We report the results in Table 6. We can find that our FRTT cannot outperform the advanced KeepTrack. We can obtain a comparable performance with TransT. In principle, our work aims to achieve joint dehazing and tracking. Our designs focus on the restoration network and joint bounding-box prediction module. Our FRTT uses a fixed template to guide tracking, which

Table 4 Performance comparisons with different haze degrees

Dataset	Degree	Metric	ToMP (Mayer et al., 2022)	PrDiMP (Danelljan et al., 2020)	DiMP (Bhat et al., 2019)	ATOM (Danelljan et al., 2019)	KeepTrack (Mayer et al., 2021)	Stark (Yan et al., 2021)	Mixformer (Cui et al., 2022b)	TransT (Chen et al., 2021)	**FRTT_C**	**FRTT_V**
Haze-OTB2015	**[50]**	**DP**	***0.167***	0.156	0.148	0.113	0.139	0.139	0.098	0.078	**0.194**	*0.183*
		AUC	***0.158***	0.144	0.137	0.103	0.129	0.151	0.118	0.087	**0.177**	*0.172*
	[100]	**DP**	***0.323***	0.271	0.277	0.265	0.290	0.234	0.214	0.280	**0.423**	*0.408*
		AUC	***0.263***	0.219	0.231	0.188	0.228	0.219	0.200	0.232	**0.375**	*0.357*
	[200]	**DP**	***0.648***	0.572	0.596	0.585	0.629	0.538	0.604	0.645	**0.717**	*0.700*
		AUC	***0.500***	0.428	0.440	0.416	0.472	0.418	0.464	0.485	**0.593**	*0.581*
	[500]	**DP**	0.821	0.790	0.774	0.755	***0.835***	0.766	***0.835***	0.814	**0.880**	*0.872*
		AUC	0.631	0.613	0.591	0.561	***0.643***	0.586	0.630	0.624	**0.669**	*0.661*
	[1000]	**DP**	0.847	0.826	0.806	0.819	***0.859***	0.810	0.831	0.839	**0.893**	*0.889*
		AUC	0.654	0.635	0.620	0.616	***0.662***	0.621	0.638	0.646	**0.678**	*0.674*
Haze-TC128	**[50]**	**DP**	***0.138***	0.128	0.126	0.117	0.115	0.082	0.078	0.069	**0.177**	*0.160*
		AUC	***0.118***	0.101	0.099	0.087	0.092	0.091	0.086	0.069	**0.163**	*0.144*
	[100]	**DP**	***0.311***	0.252	0.246	0.257	0.293	0.190	0.217	0.256	**0.399**	*0.387*
		AUC	***0.243***	0.198	0.191	0.185	0.229	0.172	0.186	0.205	**0.358**	*0.341*
	[200]	**DP**	0.561	0.533	0.539	0.544	***0.588***	0.480	0.562	0.567	*0.700*	**0.702**
		AUC	0.429	0.393	0.400	0.393	***0.441***	0.373	0.428	0.421	**0.524**	*0.517*
	[500]	**DP**	0.745	0.731	0.699	0.705	***0.775***	0.710	0.741	0.726	**0.822**	*0.811*
		AUC	0.554	0.544	0.524	0.515	***0.584***	0.530	0.553	0.544	*0.591*	**0.599**
	[1000]	**DP**	0.790	0.766	0.776	0.755	***0.813***	0.757	0.790	0.747	**0.847**	*0.832*
		AUC	0.580	0.577	0.578	0.556	***0.614***	0.562	0.590	0.562	**0.615**	*0.602*
Haze-VOT2018	**[50]**	**EAO**	0.032	0.023	0.026	0.019	***0.033***	0.004	0.003	0.010	**0.042**	*0.041*
	[100]		***0.046***	0.032	0.030	0.033	0.043	0.011	0.010	0.026	**0.083**	*0.072*
	[200]		0.087	0.100	0.101	***0.107***	0.089	0.046	0.054	0.095	**0.175**	*0.160*
	[500]		0.195	0.272	***0.296***	0.271	0.170	0.127	0.149	0.222	**0.311**	*0.306*
	[1000]		0.237	0.317	***0.358***	0.307	0.191	0.147	0.191	0.270	**0.349**	*0.342*

The top three results are highlighted in Bold, Italic and Bolditalic

Table 5 Performance comparisons with different rainy degrees

Dataset	Degree	Metric	ToMP (Mayer et al., 2022)	PrDiMP (Danelljan et al., 2020)	DiMP (Bhat et al., 2019)	ATOM (Danelljan et al., 2019)	KeepTrack (Mayer et al., 2021)	Stark (Yan et al., 2021)	Mixformer (Cui et al., 2022b)	TransT (Chen et al., 2021)	**FRTT_C**	**FRTT_V**
Rainy-OTB2015	**[100]**	**DP**	0.883	0.879	0.883	0.852	*0.894*	0.843	0.882	0.878	**0.900**	***0.893***
		AUC	0.679	0.680	0.680	0.649	***0.686***	0.649	0.681	0.649	**0.695**	*0.688*
	[200]	**DP**	***0.898***	0.874	0.880	0.852	**0.905**	0.845	0.881	0.871	*0.901*	0.890
		AUC	***0.688***	0.675	0.675	0.646	*0.692*	0.649	0.681	0.669	**0.693**	0.687
Rainy-TC128	**[100]**	**DP**	0.821	0.822	0.802	0.806	**0.856**	0.807	*0.841*	0.781	***0.823***	0.811
		AUC	0.599	*0.623*	0.603	0.599	**0.646**	0.598	***0.621***	0.581	0.600	0.599
	[200]	**DP**	0.807	***0.836***	0.792	0.785	**0.855**	0.793	*0.850*	0.779	0.825	0.809
		AUC	0.588	*0.630*	0.594	0.578	**0.642**	0.589	***0.628***	0.579	0.604	0.596
Rainy-VOT2018	**[100]**	**EAO**	0.232	*0.399*	**0.407**	***0.351***	0.204	0.159	0.196	0.286	0.327	0.333
	[200]		0.231	**0.392**	*0.384*	0.324	0.199	0.161	0.199	0.264	***0.330***	0.338

The top three results are highlighted in Bold, Italic and Bolditalic

Table 6 Tracking performance on LaSOT

	AUC	NPR	PR
KeepTrack	0.671	0.772	0.702
DiMP	0.569	0.650	0.567
TranST	0.649	0.738	0.690
FRTT_C	0.655	0.747	0.694
FRTT_V	0.651	0.745	0.690

performs inferior under natural scenarios than KeepTrack. However, our approach achieves a balanced performance between accuracy and efficiency, especially obtaining quite superior performance for dehazing tracking.

4.5 Ablation Study

To further validate the effectiveness of the proposed FRTT method, we perform an ablation study on the synthesised datasets, i.e., Haze-VOT2018 and Haze-OTB100. As reported in Table 7, the entire FRTT achieves 0.311 in terms of EAO on VOT2018, and 0.669 in terms of AUC on OTB100, with an average speed around 46 FPS. The term 'FRTT - FRT' denotes the tracking network without the restoration module, where the feature tokens X of the backbone output are directly used as the input of the transformer encoder. As the hazy visual feature is not refined to fit the target matching task, this version scarifies EAO by 0.076 and AUC by 0.072, while increasing the speed to around 65 FPS. To replace our FRT with standard multi-head self-attention layers ('FRTT - FRT + MHSA'), the performance drops 0.4% and 0.5%, demonstrating the merit of performing prompt embedded local attention.

After verifying the merit of integrating the proposed feature restoration module into the tracker, we test the effectiveness of using the transformer decoder. Specifically, 'FRTT - TDE' in Table 7 denotes the tracker without the transformer decoder (TDE), that directly predicts the corner heatmaps using the encoder output $\hat{X}_{en}$. In this version, compared to the entire FRTT, EAO is decreased from 0.311 to 0.270, and AUC is dropped from 0.669 to 0.644. Though some existing studies calculate the tracking decision with only the encoder structure, our results demonstrate the advantage of considering the decoder stage to achieve a more robust template-search fusion, compared to that in the encoder stage.

By removing the template box head from the original decoder module, the tracking performance drops from 0.311 to 0.289 in terms of EAO and from 0.669 to 0.657 in terms of AUC. The results indicate the merits of predicting template bounding box in the training stage to provide additional supervision signals for back-propagation, compared to the existing search region-only prediction paradigm. As we disregard this template bounding box head in the inference stage anyway, there is no additional computational burden for this strategy.

Table 7 The ablation study results

	Haze-VOT2018	Haze-OTB2015	
	EAO	AUC	FPS
FRTT - FRT	0.235	0.597	65
FRTT - FRT + MHA	0.307	0.664	43
FRTT	0.311	0.669	46
FRTT - TDE	0.270	0.644	50
FRTT - TBH	0.289	0.657	46
FRTT	0.311	0.669	46

5 Conclusion

To address the tracking challenges introduced by the haze imaging condition, we unified the processing of dehazing and tracking jointly at the feature level, delivering a novel haze feature restoration and tracking mechanism to enable precise object tracking. The proposed approach contains a feature restoration transformer, a template-search encoder-decider, and a joint bounding box head. The feature restoration transformer enabled feature-level recovery for the discriminative and salient visual clues. A modified transformer encoder-decoder was advocated to perform offline template-search matching learning, with both input images used in concert to supervise the target localisation. The joint bounding box head entangles the prediction clues for target localisation in both template and search region, providing accurate targetness support. The entire FRTT method achieved favourable performance in the hazy video datasets synthesised by an advanced haze generator. The collection and annotation of real hazy videos will be an integral part of future work with the aim of validating the proposed FRTT in practical scenarios. Not only single-object tracking but also multi-object tracking (Li et al., 2022; Zeng et al., 2022) will be involved in future.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (62106089, 62020106012, 62332-008, 62336004), and in part by the the Engineering and Physical Sciences Research Council (EPSRC), U.K. (Grant EP/N007743/1, Grant MURI/EPSRC/DSTL, and Grant EP/R018456/1).

Availability of Data and Materials The datasets supporting the conclusions of this article are included within the article.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Berman, D., & Avidan, S. (2016). Non-local image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1674–1682).

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. In *Proceedings of the European conference on computer vision* (pp. 850–865).

Bhat, G., Danelljan, M., Gool, L. V., & Timofte, R. (2019). Learning discriminative model prediction for tracking. In *IEEE international conference on computer vision* (pp. 6182–6191).

Cai, B., Xu, X., Jia, K., et al. (2016). Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing, 25*(11), 5187–5198.

Cao, Z., Fu, C., Ye, J., Li, B., & Li, Y. (2021). Hift: Hierarchical feature transformer for aerial tracking. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15457–15466).

Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., & Fu, C. (2022). Tctrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14798–14808)

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision* (pp. 213–229).

Charbonnier, P., Blanc-Feraud, L., Aubert, G., & Barlaud, M. (1994). Two deterministic half-quadratic regularization algorithms for computed imaging. *IEEE International Conference on Image Processing, 2*, 168–172.

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021). Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8126–8135).

Chen, Z., Zhong, B., Li, G., Zhang, S., & Ji, R. (2020). Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6668–6677).

Cui, Y., Jiang, C., Wang, L., & Wu, G. (2022a). Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13608–13618).

Cui, Y., Jiang, C., Wang, L., & Wu, G. M. (2022b). End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18–24).

Danelljan, M., Häger, G., Khan, F. S., & Felsberg, M. (2017). Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(8), 1561–1575.

Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2019). Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Danelljan, M., Gool, L. V., & Timofte, R. (2020). Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7183–7192)

Dosovitskiy, A., Beyer, L., Kolesnikov A, Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., & Uszkoreit, J. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., & Ling, H. (2019). Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5374–5383).

Gao, S., Zhou, C., Ma, C., Wang, X., & Yuan, J. (2022). Aiatrack: Attention in attention for transformer visual tracking. In *European conference on computer vision*, Springer (pp. 146–164).

Guo, D., Wang, J., Cui, Y., Wang, Z., & Chen, S. (2020). Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6269–6277).

Guo, D., Shao, Y., Cui, Y., Wang, Z., Zhang, L., & Shen, C. (2021). Graph attention tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9543–9552).

He, K., Sun, J., & Tang, X. (2010). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(12), 2341–2353.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–778).

Held, D., Thrun, S., & Savarese, S. (2016). Learning to track at 100 fps with deep regression networks. In *Proceedings of the European conference on computer vision*, Springer (pp. 749–765).

Hong, M., Xie, Y., Li, C., & Qu, Y. (2020). Distilling image dehazing with heterogeneous task imitation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3462–3471).

Huang, L., Zhao, X., & Huang, K. (2019). Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*, 1562.

Jiao, L., Wang, D., Bai, Y., Chen, P., & Liu, F. (2021). Deep learning in visual tracking: A review. *IEEE Transactions on Neural Networks and Learning Systems, 34*, 5497.

Kang, B., Chen, X., Wang, D., Peng, H., & Lu, H. (2023). Exploring lightweight hierarchical vision transformers for efficient visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9612–9621)

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Cehovin Zajc, L., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., & Fernandez, G. (2018). The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European conference on computer vision (ECCV) workshops* pp 1–50.

Li, A., Lin, M., Wu, Y., Yang, M. H., & Yan, S. (2016). Nus-pro: A new visual tracking challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 335–349.

Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp 8971–8980).

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J,. & Yan, J. (2019). Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4282–4291).

Li, S., Danelljan, M., Ding, H., Huang, T. E., & Yu, F. (2022). Tracking every thing in the wild. In *European conference on computer vision*, Springer (pp. 498–515).

Li, S., Yang, Y., Zeng, D., & Wang, X. (2023). Adaptive and background-aware vision transformer for real-time UAV tracking. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13989–14000).

Liang, P., Blasch, E., & Ling, H. (2015). Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Transactions on Image Processing, 24*(12), 5630–5644.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll'ar, P., & Zitnick, C. L. (2014) Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision* (pp 740–755).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101

Mayer, C., Danelljan, M., Paudel, D. P., & Van Gool, L. (2021). Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13444–13454).

Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D. P., Yu, F., & Van Gool, L. (2022). Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8731–8740).

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., & Ghanem, B. (2018). Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision* (pp. 300–317).

Qin, X., Wang, Z., Bai, Y., Xie, X., & Jia, H. (2020). Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI conference on artificial intelligence* (pp. 11908–11915).

Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., & Yang, M. H. (2016). Single image dehazing via multi-scale convolutional neural networks. In *Proceedings of the European conference on computer vision* (pp. 154–169).

Song, Y., He, Z., Qian, H., & Du, X. (2023). Vision transformers for single image dehazing. *IEEE Transactions on Image Processing, 32*, 1927–1941.

Tao, R., Gavves, E., & Smeulders, A. W. (2016). Siamese instance search for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, IEEE (pp. 1420–1429).

Wang, N., Shi, J., Yeung, D. Y., & Jia, J. (2015). Understanding and diagnosing visual tracking systems. In *IEEE international conference on computer vision* (pp. 3101–3109).

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. H. (2019). Fast online object tracking and segmentation: A unifying approach. In *IEEE conference on computer vision and pattern recognition* (pp. 1328–1338).

Wang, X., Tang, J., Luo, B., Wang, Y., Tian, Y., & Wu, F. (2021). Tracking by joint local and global search: A target-aware attention-based approach. *IEEE Transactions on Neural Networks and Learning Systems, 33*(11), 6931–6945.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22–31).

Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In *IEEE Conference on computer vision and pattern recognition* (pp. 2411–2418).

Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1834–1848.

Xu, T., Feng, Z. H., Wu, X. J., & Kittler, J. (2019). Joint group feature selection and discriminative filter learning for robust visual object tracking. In *Proceedings of the IEEE international conference on computer vision* (pp. 7950–7960).

Xu, T., Feng, Z. H., Wu, X. J., & Kittler, J. (2020). Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(10), 3727–3739.

Xu, T., Feng, Z., Wu, X. J., & Kittler, J. (2021). Adaptive channel selection for robust visual object tracking with discriminative correlation filters. *International Journal of Computer Vision, 129*, 1359–1375.

Xu, T., Feng, Z., Wu, X. J., & Kittler, J. (2023). Toward robust visual object tracking with independent target-agnostic detection and effective siamese cross-task interaction. *IEEE Transactions on Image Processing, 32*, 1541–1554.

Xu, Y., Wang, Z., Li, Z., Yuan, Y., & Yu, G. (2020b). SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI conference on artificial intelligence* (pp. 12549–12556).

Yan, B., Peng, H., Fu, J., Wang, D., & Lu, H. (2021). Learning spatio-temporal transformer for visual tracking. arXiv preprint arXiv:2103.17154

Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., & Wei, Y. (2022). Motr: End-to-end multiple-object tracking with transformer. In *European conference on computer vision*, Springer (pp. 659–675).

Zhang, Y., Ding, L., & Sharma, G. (2017). Hazerd: An outdoor scene dataset and benchmark for single image dehazing. In *ICIP* (pp. 3205–3209).

Zhu, H., Peng, X., Chandrasekhar, V., Li, L., & Lim, J. H. (2018). Dehazegan: When image dehazing meets differential programming. In *International joint conference on artificial intelligence* (pp. 1234–1240).

Zhu, Q., Mai, J., & Shao, L. (2014). Single image dehazing using color attenuation prior. In *BMVC*.